オンデマンド情報抽出

関根聡
ニューヨーク大学

1. はじめに

世の中の多くの情報はテキストの形式で生産され、伝達され、消費されている。そういったテキストの中には、本質的に構造をもつ情報が多く含まれている。例えば、繰り返し報道される人事情報は、ある企業におけるある人のある役職への昇格や就任という同じタイプの構造を持った情報である。情報抽出とは、構造化されていないテキストからこのようなある特定の構造を持った情報を抽出する技術である。しかし、情報抽出の技術には大きな壁があった。それは、特定の情報を抽出するための知識をどのように獲得するかという点である。特定の情報を抽出するための知識とは究極的には表現と情報のリンクであり、例えば表現パターンや辞書のような形式で実現される。1994 年頃まで行なわれていた MUC では、トピックが与えられた後に 1 ヶ月の準備期間があり、その間に表現パターンなどの知識の作成が主に人手で行なわれていた。この知識作成の困難さに伴うポータビリティの問題は情報抽出において非常に重要な問題であり大きな障害であった。この問題を何らかの形で解決することによって、情報抽出を容易に使える真のアプリケーションにすることができると考えている。

この問題に対し、我々は「オンデマンド情報抽出」という概念を打ち立てた(Sekine 06)。それは、ユーザーが指定したトピックに対し、自動的にそのトピックに内在する情報の構造を認識し、その情報のインスタンスを抽出し、その結果をテーブルの形式で表示するというものである。この新しい情報抽出のパラダイムは、自然言語の最先端の研究成果を利用して実現可能になった。それは、一般的な自然言語ツールや拡張固有表現の整備などに加えて、教師なし学習による知識獲得が重要な技術となっている。本デモンストレーションでは、この新しいパラダイムにあるオンデマンド情報抽出システムの紹介を行なう。

2. 概要

開発したオンデマンド情報抽出システムは、ユーザーの知りたいトピックについてのキーワード（例えば「企業買収」「M&A」）を入力とし、その約 1 分後に、その情報に関するテーブルを出力する（図 2~5)。システム構成図を図 1 に示し、これを基にシステムの動作を説明する。システムには 6 つのコンポーネントがあり、その間をデータが流れていく。なお、システムの対象言語は英語であり、コーパスは 11 年分の新聞記事（APW,NYT 各 3 年分、英字 XINHA の 5 年分）を使用している。

１） 情報検索

与えられた検索キーワードに対して、コーパスから関連記事を検索する。ここでは、簡単な TF/IDF によるシステムを開発し、使用している。

２） パターンの獲得

コーパス中のテキストは予め、POS タガー、係り受け解析、拡張固有表現タガー(本節 5 を参照のこと)で解析されている。情報検索の結果集められたテキストから以下の条件に合うすべての部分的な係り受け構造を抽出する。条件は 2～6 ノードであること、ヘッドに述語か名詞化された述語がくること、少なくとも一つの固有表現を含むこと、その頻度が予め設定する頻度の範囲内であること、である。最後の条件は、あまり頻度が大きくて一般的過ぎるものや、頻度が少なすぎてノイズの可能性があるものを除くための処置である。このようにして抽出された部分係り受け構造は、検索されたテキスト内で比較的多く現れるものに高いスコアを与え、スコアの高い特定数のものをこのトピックのパターンとし、後のプロセスで使用する（Sudo et al. 03)。高速な実行のために、パターンとなり得る部分的係り受け構造の抽出やドキュメントのリンク、全体における頻度などは予めオフラインの形で計算しておく。

３） パラフレーズの獲得

抽出したパターン同士の意味的な関係、つまりパラフレーズの関係を認識する。これは、異なる表記で表現されているため、異なるパターンにマッチした内容であるが、同一種類のイベントの内容として、一つのテーブル内に表示するために必要である。このシステムでは同義語辞書である WordNet と、自動的なパラフレーズ発見システム(Hasegwa et al. 04) (Sekine 05)を使用している。この知識も予めオフラインで収集しておく。

４） テーブル作成

すでにパラフレーズでリンクされたパターンのセットは獲得している。これらのパターンを実際のテキストにマッチングさせ情報を抽出する。対象のテキストは１で検索したテキストであり、抽出する情報は全て拡張固有表現のいずれかのカテゴリとしてタグ付けされたものである。

５） 自然言語処理ツール

自然言語処理ツールとしては、自ら開発した POS タガー、係り受け解析を使用している(OAK HP)。

６） 拡張固有表現タガー

情報抽出の対象となるものは固有表現であることが多い。しかしながら、一般的に定義されている固有表現では、任意の情報要求に対して満足できるカテゴリを提供していない。例えば、よく用いられる MUC の 7 種類や IREX の 8 種類の固有表現（人名、地名、組織名、固有物名、時間、日時、割合、金額）では、イベント名、病気名、多くの数値表現など多くの種類の重要な固有表現が含まれていない。我々は幅広い固有表現に対応するように 140 種類～210 種類にわたる拡張固有表現を設計し、タガーを作成している(Sekine et al. 04)(ENE HP)。

3. 高速化のための手段

本システムの実現における問題点の一つにスピードがある。本システムでは、キーワードで検索した約 200 のテキストに含まれる数十万規模の部分係り受け構造から、もっともスコアの高い 1000 のパターンを抽出しているが、最新のデータマイニングの技術を使用しても、この作業をオンラインで行なった場合、14 分程度かかる。これは検索結果のテキストが動的に決まることに起因している。この問題を回避するために、我々は、11 年分の新聞記事に含まれる部分係り受け構造のすべてを予め作成し、それがどのテキストに含まれていたか、全体での頻度はいくつであったのかを計算した。11 年分のコーパスには、110 万記事、2500 万文が含まれており、そこから頻度 1 以外で、条件を満たす部分係り受け構造の異なり数は 3900 万に及ぶ。その内、頻度が 11 以上 10,000 以下のものは約 98 万であり、この全ての部分係り受け構造と、文、記事に ID を付けそれらをリンクさせてデータベース化している。現実のシステムは全てこれらのデータベース上の操作で実装されている。

4. システム動作シナリオ

例を用いてシステムの動作を説明する。図２は開始画面である。ここにあるテキストボックスにユーザーの情報要求をキーワードで入力する。このキーワードは任意ではあるが、最初の情報検索のコンポーネントによって適切なテキストが検索されるようなキーワードか好ましい。「企業の M&A」といった抽象的な表現では、実際の企業買収の記事は検索されないことが多いため、「買収」「合併」「株式公開買付」といった該当記事によく現れる単語を入力する。キーワード入力後 1 分程度で図３にあるような表が出力される。抽出できた全ての表が表示され、文 ID、記事 ID と共に、情報の内容が表示される。もし、該当情報が見つからなかった場合には、その部分は空欄で示される。また、図４にある該当記事や、図５にある対応パターンなどの表示も可能である。

5. 評価

システムの評価を、米国で行なわれている情報抽出のプロジェクトである ACE のイベントタイプの内の 20 トピックを使って行なった。イベントは、例えば「企業合併」「就任」「逮捕」「選挙」「罰金」「刑の執行」などである。抽出した表に対する主観的な評価では、2 トピックがその表示だけで情報要求が満足できるほど有効、12 トピックが情報検索の記事を並べられるよりも役に立つ、6 トピックが役に立たないという結果であった。また、情報の内容の正確性では、ランダムに抽出した 100 行の内、4 行で一部に間違いがあり、14 行が完全に間違っていた。間違いの大きな理由は固有表現抽出の誤りであった。また、語義の曖昧性や情報検索の誤りによる間違いも見られた。

6. まとめ

オンデマンド情報抽出という新しいパラダイムの情報抽出を提案し、そのシステムを開発しデモを行なう。改良点としては、まず基礎的なコンポーネントの精度向上が挙げられる。拡張固有表現、係り受け解析、パターン抽出、パラフレーズの発見だけではなく、照応解析が非常に重要であろうことを認識している。また、出力した表において、タイトル行がそのイベントにおけるエンティテーの役割ではなく、固有表現の種類名しか載せられていないのは改良する必要があるだろう。システムへの入力がテキストに現れているキーワードでなくてはならないというのも厳しい制約であり、できれば解消していきたいと考えている。

謝辞

This research was supported in part by the National Science Foundation under Grant IIS-00325657. This paper does not necessarily reflect the position of the U.S. Government. また、研究に有益な助言などを与えてくれたり、部分的な開発を行なっていただいた、Grishman 教授、野畑氏、須藤氏、新山氏、竹内氏、長谷川氏、Marx 氏、村上氏、Young 氏、Reeves 氏、Stenchikova 氏、尾田氏に感謝する。

参考文献

ENE homepage: http://nlp.cs.nyu.edu/ene

OAK homepage: http://nlp.cs.nyu.edu/oak

Takaaki Hasegawa, Satoshi Sekine and Ralph Grish-man 2004. “Discovering Relations among Named Entities from Large Corpora”, ACL-04

Satoshi Sekine. “On-Demand Information Extraction”, ACL-COLING-2006.

Satoshi Sekine, C. Nobata, “Definition, Dictionary and Tagger for Extended Named Entities”. LREC 2004

Satoshi Sekine. 2005. “Automatic Paraphrase Discov-ery based on Context and Keywords between NE Pairs”. IWP-05

Kiyoshi Sudo, Satsohi Sekine and Ralph Grishman. 2003. “An Improved Extraction Pattern Representa-tion Model for Automatic IE Pattern Acquisition”. ACL-03.

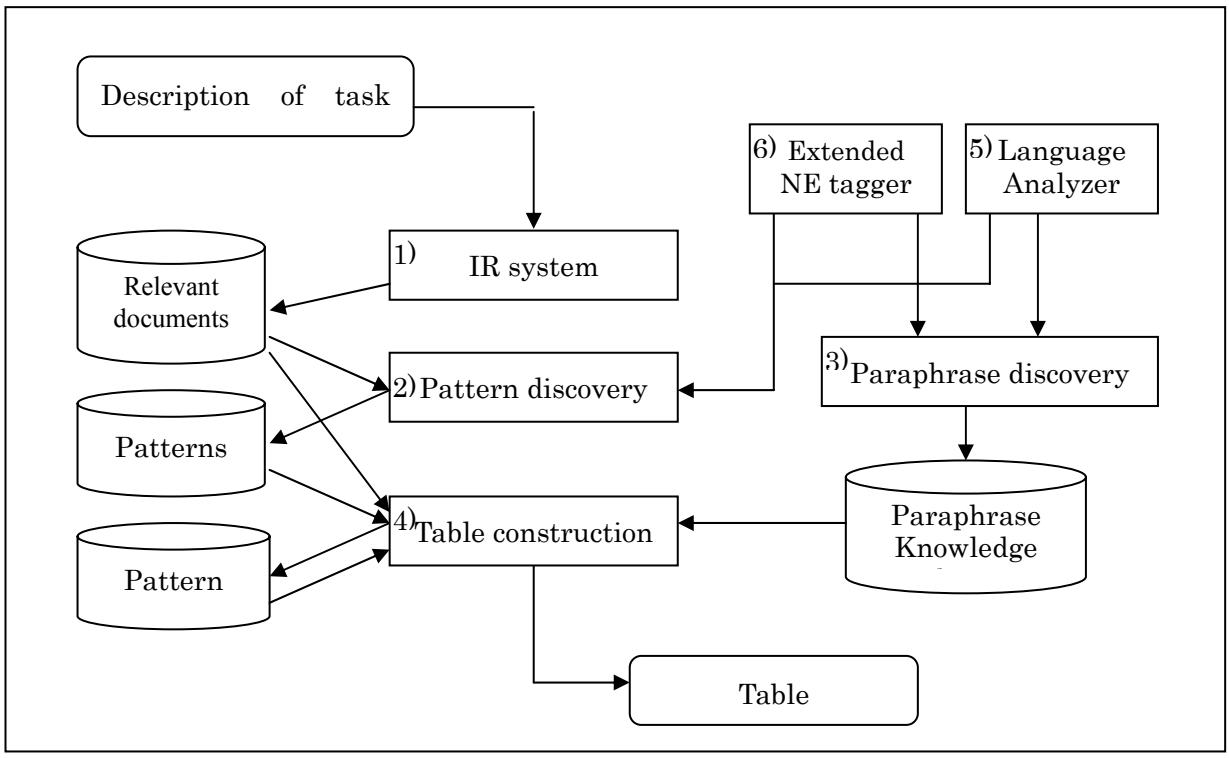
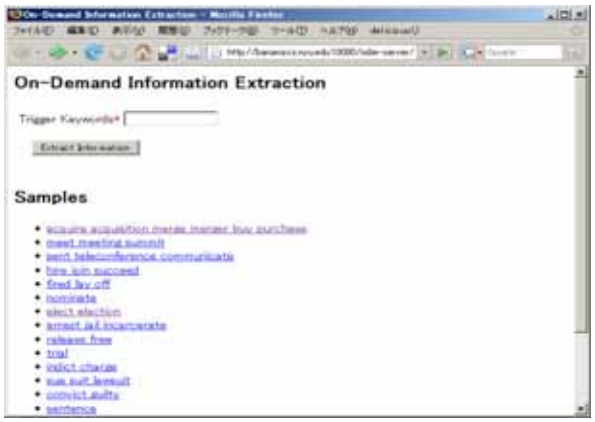


図 1：システム構成図



左上）図２　開始画面
中央）図３　結果表示画面
左下）図４　パターン表示画面
右下）図５　該当テキスト表示画面